Procedia Computer Science

Volume 80, 2016, Pages 2276–2280

ICCS 2016. The International Conference on Computational Science

BEAM: A computational workflow system for managing and modeling material characterization data in HPC environments

E. J. Lingerfelt[1], A. Belianinov[1], E. Endeve[1], O. Ovchinnikov[2], S. Somnath[1], J. M. Borreguero[1], N. Grodowitz[1], B. Park[1], R. K. Archibald[1], C. T. Symons[1], S. V. Kalinin[1], O. E. B. Messer[1], M. Shankar[1], and S. Jesse[1]

[1]*Oak Ridge National Laboratory, Oak Ridge, TN*

lingerfeltej@ornl.gov, belianinova@ornl.gov, endevee@ornl.gov, somnaths@ornl.gov, borreguerojm@ornl.gov, grodowitznt@ornl.gov, parkbh@ornl.gov, archibaldrk@ornl.gov, symonsct@ornl.gov, sergei2@ornl.gov, bronson@ornl.gov, shankarm@ornl.gov, sjesse@ornl.gov

[2]*Vanderbilt University, Nashville, TN*

oleg.ovchinnikov@vanderbilt.edu

Abstract

Improvements in scientific instrumentation allow imaging at mesoscopic to atomic length scales, many spectroscopic modes, and now—with the rise of multimodal acquisition systems and the associated processing capability—the era of multidimensional, informationally dense data sets has arrived. Technical issues in these combinatorial scientific fields are exacerbated by computational challenges best summarized as a necessity for drastic improvement in the capability to transfer, store, and analyze large volumes of data. The Bellerophon Environment for Analysis of Materials (BEAM) platform provides material scientists the capability to directly leverage the integrated computational and analytical power of High Performance Computing (HPC) to perform scalable data analysis and simulation via an intuitive, cross-platform client user interface. This framework delivers authenticated, "push-button" execution of complex user workflows that deploy data analysis algorithms and computational simulations utilizing the converged compute-and-data infrastructure at Oak Ridge National Laboratory's (ORNL) Compute and Data Environment for Science (CADES) and HPC environments like Titan at the Oak Ridge Leadership Computing Facility (OLCF). In this work we address the underlying HPC needs for characterization in the material science community, elaborate how BEAM's design and infrastructure tackle those needs, and present a small sub-set of user cases where scientists utilized BEAM across a broad range of analytical techniques and analysis modes.

Keywords: Computational workflows; HPC workflows; data management; materials science; materials modeling; scalable data analysis; user experience design; multi-tier architectures

 Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2016

doi:10.1016/j.procs.2016.05.410

1 Introduction

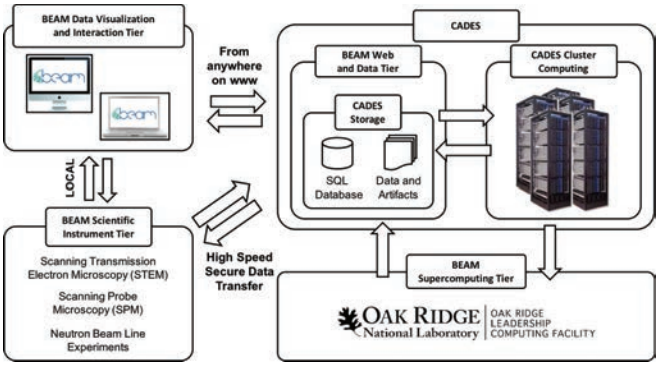
Scanning (transmission) electron microscopy (S(T)EM) and scanning probe microscopy (SPM) along with their associated spectroscopies are well established, robust imaging tools that have proved to be cornerstones in the visualization of structure and functionality of materials with atomic resolution. The ultimate goal of these tools is to observe and quantitatively correlate structure-property relationships with functionality—by evaluating chemical, electronic, optical, and mechanical properties of individual atomic and nanometer-sized structural elements. In essence, these microscopes are the final frontiers for materials science, as they provide control and visualization of matter at the atomic scales. The natural evolution of instrument hardware and data processing technologies has allowed determination of atomic positions with sub-10pm precision, and enabled visualization of previously elusive properties. Ideally, information should be collected as a function of global stimuli, such as temperature or uniform electric field, as well as local stimuli induced by additional probe or ionic interactions. However, meeting this scientific challenge requires acquiring orders of magnitude more data and extracting a hidden wealth of information necessitating a drastic improvement in *the capability to transfer, store, and analyze multidimensional data sets.*

On the other end of the instrumentation spectra lie massive beamline facilities that require implementation of data provenance strategies well before the construction of the instrument begins. While their approach is much more robust in nature, significant obstacles also related to data transfer, visualization, and modeling are regularly encountered by scientific staff at facilities such as the Spallation Neutron Source (SNS) at Oak Ridge National Laboratory (ORNL), echoing a similar message of their smaller scale imaging instrumentation colleagues. The similarities in these challenges, across what seem to be drastically different types of instrumentation, could perhaps be understood better from a vantage point of how the data itself is generated. The digital age brought about miniaturized detector technologies that begin to rival the pixel densities of measurements made at beam lines. Compounded by the stability of modern imaging platforms, fast electronics, and generally much higher veracity signal, a desktop machine can now produce similar amounts of data as detectors at the Large Hadron Collider.

2 Bellerophon Environment for Analysis of Materials (BEAM)

Typical scientific data problems involve large sets of data slowly processed with individually purchased licensed software packages. Any workflow system seeking to solve these issues must provide near real-time analysis via modern algorithms at multiple scales accessible to anyone regardless of skillset and depth of computational background. Users should be able to effortlessly utilize HPC compute and data resources with no previous experience as a computer programmer or an HPC user. The BEAM software system provides users a secure, flexible environment for analysis and modeling of material characterization data in HPC environments through a central, highly intuitive portal. It enables users to easily configure and dynamically generate complex data analysis and simulation workflows designed for a variety of HPC platforms. Once the user selects the desired HPC resource and number of CPU cores (and enters her authentication information, if required), she is able to deploy and execute the resulting workflow with the push of a button. BEAM's user interface then provides real-time monitoring of the workflow and the ability to override the execution remotely. Once the workflow has completed, users can visualize and export the results.

BEAM's unique infrastructure, which extends the multi-tier architecture of Bellerophon[1], offers a mechanism for efficient delivery of scalable algorithms developed by mathematicians and computational scientists, as well as authenticated access to data analysis services and remote data stewardship from anywhere at any time. This architecture is realized by a system of distributed software components and services that connect locally to scientific instruments and intercommunicate over the network as illustrated in Figure 1. It combines scientific instruments with web and data



services and HPC resources via a lightweight user interface for data visualization and interaction. The Compute And Data Environment for Science (CADES) at ORNL provides BEAM with scalable computing and software support, as well as high performance data storage.

Figure 1: A schematic representation of BEAM's multi-tier architecture, which interconnects scientific instruments and HPC resources.

3 Use Scenarios

In this section, we illustrate scenarios already implemented in BEAM to aid in data processing, analysis, and visualization. We chose examples to illustrate globally common capabilities to the area of material science, which include curve fitting, multivariate analysis, and near real-time theory-in-the-loop. A more diverse set of algorithms, integration with more instrument platforms, and a wider array of available options for sharing and visualization will follow.

Band Excitation Atomic Force Microscopy (BE-AFM)[2] is a powerful group of imaging techniques that relies on collecting broadband AFM-cantilever response around a resonance. The basic data unit for BE is a 3D dataset where a frequency band vector of the cantilever transfer function is captured at each spatial pixel. More complex applications of BE vary parameters such as AC and DC bias, temperature, probe offset, laser position, and many others. This results in large multidimensional datasets that are far beyond the timely processing capability of single workstations. Processing is made more difficult because the basic data unit must be fit to a parametric model. In order to reduce this parameterization time, the fitting was re-implemented in BEAM via a Levenberg-Marquardt algorithm with least squares minimization using a Fortran implementation in parallel with MPI. When deployed on a HPC resource, a speed up of 800x to 2800x was immediately achieved, scaling linearly with resource consumption. Furthermore, BEAM's BE Analyzer software tool enabled researchers to securely upload and access data in their private data storage area at CADES; view an interactive surface of spectrally averaged data, Figure 2a; easily execute the new fitting implementation with HPC resources; and graphically analyze and filter the results of the fitting using customizable 2D color maps of the fitted parameters.

S(T)EM is a well-established, robust set of imaging and characterization tools that formed the bedrock of atomic resolution studies of materials over the last sixty years. Electron microscopy is currently transitioning from the traditional model of producing a few images, through the current state where access to image and spectral information is significantly increased, to a new state where a very large volume of data (movies and multi-dimensional series) can be rapidly obtained. Data rates of 4 Gb/sec for emerging imaging technologies using standard instrument parameters are easily accessible.

Alternatively, mining already existing images beyond simple qualitative analysis is an attractive problem since so much data has been collected over the lifetime of the technique that has yet to be fully analyzed. Specifically, identifying repeated, statistically defined units based on graph partitioning of the underlying atomic lattice[3]; creating a basis for development of image genomes; and furthering development in structure-property correlative libraries, are particularly appealing avenues since the most effective pathway to achieving these goals lies through HPC systems. The BEAM framework contains a wide (and growing) number of image-processing algorithms revolving around

multivariate analysis and clustering specifically designed to extract quantitative information from otherwise qualitatively processed data. Figure 2b illustrates local crystallography algorithms used to quantify lattice structure on a STEM image that has been entirely pre-processed, filtered, and analyzed using an HPC resource. While seemingly trivial, collecting such an image takes milliseconds thus placing a premium on analysis speed. Access to near real-time capabilities is even more important as in the cases of movies, and imaging as a function of other parameters described above.

The Spallation Neutron Source (SNS) at ORNL is the most intense pulsed neutron beam system in the world, designed to attract the most innovative scientific and industrial work. When the neutron beam is impinged onto a sample the neutron particles scatter. The energy and angle at which this process occurs reveals the molecular and magnetic structure in materials, like high-temperature superconductors, polymers, metals, and a broad range of soft materials. Computational workflows developed at the Center for Accelerating Materials Modeling (CAMM) at ORNL allow neutron scientists to directly integrate beamline results into Molecular Dynamics (MD) simulations to guide their subsequent experiments. In this theory-in-the-loop use case, workflows of parameter refinement of force fields used in MD simulations by iterative optimization against Quasi-Elastic Neutron Scattering (QENS) data taken at the BASIS beam line of the SNS were examined[4]. A multi-stage workflow utilizing parallel processing was designed and implemented on a cluster compute resource at CADES to simulate the physical system and compare the results to the experimental data. The workflow chained computation using NAMD[5], to simulate single molecule behavior; Amber Tools[6], to the remove global rotations and translations; Sassena[7], to calculate structure factors; and Mantid[8], to compare the simulation results directly against experimental data. The final optimization in BEAM guides the user to rapidly calculate neutron scattering function from their simulation, modifies force-field parameters to achieve optimal approximation against experiment, and visualizes the results with an interactive multidimensional data view of simulation results versus experimental neutron data as shown in Figure 2c.

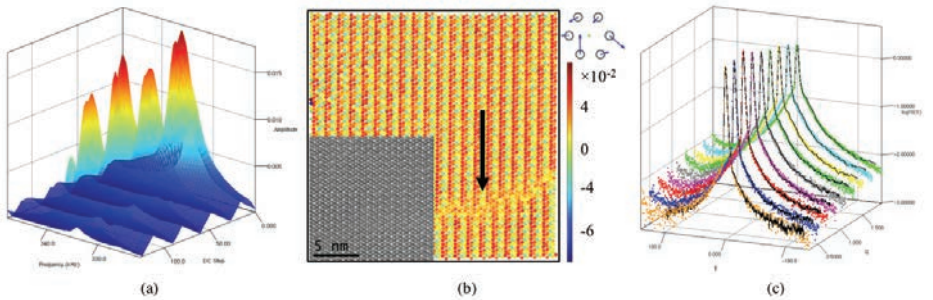


Figure 1: (a) An interactive surface rendering of spectrally averaged BE data. (b) Atomically resolved STEM image of a single M2 phase of Mo-V-Te oxide., Result of PCA (Eigenvector shown as a lattice structural mode in the upper right) of found atomic positions for 6-member neighborhood. Black line indicates a lattice shift invisible in the original image. (c) An interactive 3D plot of a guided force-field parameter simulation calculating the optimization of potential energy barrier to methyl rotations in octamethyl silsesqioxane.

4 Discussion

Deep data discovery in materials only begins with the data obtained from advanced microscopes, detectors, and neutron beams. The ability to integrate these instruments with modern HPC platforms—and the attendant expertise of computational professionals, mathematicians, visualization specialists, and storage experts—is a powerful method to advance discovery from experiment. BEAM attempts to reduce the barrier between experimental data and computational resources to enable characterization aspects of material science that have heretofore been carried out on more modest computational resources. Accelerating these characterization tasks via HPC increases the rate at which sophisticated analyses can be brought to bear on experimental data. This rapid data-analysis cycle provides for a

strong interaction of experiment and theory. As an example, the multimodal, hyperspectral data collected via next-generation microscopy techniques are an amalgam of what is typically independently processed by well-established, theoretical techniques that utilize self-contained approaches. These techniques, however, are rarely cross-validated. Data storage, rapid pre-processing, and the ability to apply multiple theoretical treatments can intertwine experiment and theory into a single, streamlined analysis process. It is precisely this sort of interaction that can be enabled by BEAM in an HPC environment.

Emerging trends in the characterization of materials places heavy emphasis on combinatorial imaging that correlates spatial, chemical, and physical information. Serious challenges in processing these data have been slowly, but steadily, addressed by the scientific community on many fronts. However, scaling, validating, and cross-correlating these independent efforts has been a challenge. BEAM directly addresses these challenges by providing an environment where all of these tasks can be performed in a seamless manner using both state-of-the-art experimental facilities and best-in-class HPC resources.

Acknowledgements

This work is partially supported by the Laboratory Directed Research and Development (LDRD) program at ORNL, which is managed by UT-Battelle, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC05-00OR22725 (E.J.L., A.B., E.E., O.O., S.S., C.T.S., S.V.K., M.S., and S.J.). This research was conducted at the Center for Nanophase Materials Sciences and the Spallation Neutron Source, which are DOE Office of Science User Facilities. Research by J.M.B. is supported by the Center for Accelerating Materials Modeling (CAMM), which is funded by DOE Basic Energy Sciences under FWP-3ERKCSNL. This research used resources of ORNL's Compute and Data Environment for Science (CADES) and the Oak Ridge Leadership Computing Facility (OLCF), which are supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. The mathematical aspects were sponsored by the applied mathematics program at the DOE by the ACUMEN project. We would like to also acknowledge ORNL’s John Quigley, Ken Barker, and Matt Disney for their assistance integrating the BEAM software system with CADES.

References

1 Lingerfelt, E. J., *et al.* Near Real-time Data Analysis of Core-collapse Supernova Simulations with Bellerophon. *Procedia Computer Science* **29**, 1504-1514 (2014).

2 Jesse, S. *et al.* Band excitation in scanning probe microscopy: recognition and functional imaging. *Annual review of physical chemistry* **65**, 519-536 (2014).

3 Belianinov, A. *et al.* Identification of phases, symmetries and defects through local crystallography. *Nature communications* **6** (2015).

4 Borreguero, J. M. & Lynch, V. E. Molecular dynamics force-field refinement against quasi-elastic neutron scattering data. *J. Chem. Theory Comput.*, 2016, **12** (1), pp 9–17.

5 Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *Journal of computational chemistry* **26**, 1781-1802 (2005).

6 Salomon-Ferrer, R. *et al.* An overview of the Amber biomolecular simulation package. *WIREs Comput. Mol. Sci.* **3**, 198-210 (2013).

7 Lindner, B. & Smith, J. C. Sassena—X-ray and neutron scattering calculated from molecular dynamics trajectories using massively parallel computers. *Computer Physics Communications* **183**, 1491-1501 (2012).

8 Arnold, O. *et al.* Mantid—Data analysis and visualization package for neutron scattering and SR experiments. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **764**, 156-166, (2014).